An Anatomy of Security Conversations in Stack Overflow

Conference or Workshop Item

An Anatomy of Security Conversations in Stack Overflow

Tamara Lopez*, Thein Tun*, Arosha Bandara*, Mark Levine†, Bashar Nuseibeh*‡ and Helen Sharp*
*School of Computing & Communications, The Open University, Milton Keynes, UK
†Department of Psychology, University of Exeter, Exeter, UK
‡Lero - The Irish Software Research Centre, University of Limerick, Limerick, Ireland
Email: *firstname.lastname@open.ac.uk, †firstname.lastname@exeter.ac.uk, ‡firstname.lastname@lero.ie

***Abstract*—As software-intensive digital systems become an integral part of modern life, ensuring that these systems are developed to satisfy security and privacy requirements is an increasingly important societal concern. This paper examines how secure coding practice is supported on Stack Overflow. Although there are indications that on-line environments are not robust or accurate sources of security information, they are used by large numbers of developers. Findings demonstrate that developers use conversation within the site to actively connect with and tend to security problems, fostering knowledge, exchanging information and providing assistance to one another.**

***Index Terms*—secure software development, collaborative environments, empirical studies**

I. Introduction

The pervasive adoption of digital technologies across many aspects of daily life means that software-intensive systems are an integral part of how we live. This has extended the social obligation of governments to provide security for their citizens to include cybersecurity [1], making this a key area of concern for modern society. There is a growing list of cybersecurity tools, guidance, training materials and case studies, yet the number of breaches seems to be continuing. Indeed many recent breaches, such as those experienced by Equifax [2] and Illinois State Board of Elections [3], exploit known vulnerabilities in software systems.

So what is going on? Is the problem that developers don't know enough, or that they don't care, or that programming languages lack suitable security features? Different theories exist, but in other contexts, developers *do* ask each other for help and learn from each other such that knowledge grows within the community.

Security is, in part, a social phenomenon. Workers bring to the desk a degree of awareness about security formed on the job and in wider engagement in the world [4]. They exhibit a sense of responsibility toward security and their organizations [5]. The ideal within organisations is to achieve a "security culture", in which behaving securely is an implicit part of behavior [4]. A range of voices and skills contribute to this process: individuals with different levels of commitment to being secure [6], including those that have basic security awareness, and those who are fully committed, security "champions" [7].

These views on security lead to questions about what security is within software development, and in the context of this study, within Stack Overflow. Do developers who participate on the site view security as a duty, or something different? Are values associated with secure coding practice? Is security something to comply with, or to champion?

This study looks at how developers talk to each other about security in Stack Overflow, and hence how understanding of security and secure practices is developed and disseminated among practitioners in this on-line environment. It is part of a larger program of research that is investigating ways to initiate and sustain secure software culture using established frameworks of personal motivation and team culture [8], [9].

Within the program, this is the second study examining Stack Overflow. The first study examined how developers talk to one another in a set of comment streams for questions given the "security" tag in Stack Overflow. The prior analysis suggested that talk about security within Stack Overflow includes information about technical solutions to programming problems, but also statements about personal values and attitudes like responsibility, trust, and fear [10].

This report examines interactions within Stack Overflow accepted answer comment streams for the same set of data. Taking an ethnographic approach [11], the study asks:

How do developers on Stack Overflow engage with one another when dealing with issues related to security?

II. Background

Stack Overflow is a question and answer site in which developers can ask questions about programming problems they are solving, and get answers. One of several Q&A sites within the Stack Exchange family, the site was founded in 2008 by Jeff Atwood, who compared the site to other websites that invite public participation, noting that it is a resource "by programmers, for programmers." [12].

A social learning environment [13], the site is part of a new wave of social media that have given rise to the social programmer [14]. Recent studies within software engineering have examined individual Stack Overflow channels, examining how knowledge is shared and formed within the R channel [15], and finding evidence for differences in use between this environment and other communication channels such as

mailing lists [16]. In a qualitative analysis, Nasehi et al. asked what makes a good question [17].

Among on-line sources, Stack Overflow is reported to be the most popular source for learning to code, even among developers who have computer science degrees[1]. In 2017, 90% of Stack Overflow survey respondents reported finding an answer to solve a coding problem on the site[2]. In the survey from 2018, almost 60% of respondents identify as back-end developers and 81% of the professional developers have coding as a hobby. Also, 87,450 respondents out of 98855 are professionals. Roughly two thirds of survey respondents reported that they visit the site at least once a day.

However, the numbers of active participants is much smaller: slightly more than half report participating in streams less than once per month or not at all. Characterised within software engineering research as "one-day flies", possible reasons for the lack of ongoing activity in this user group may be due to the quality of their original post, negative feedback they received in response to it, or efforts to "game" their reputation on the site [18]. However, it is also possible that the database has grown so large that many users are able to find answers to their questions on the site without posting, suggesting that the user community of Stack Overflow contains a number of "legitimate peripheral", rather than active, participants [19]. A final explanation for peripheral participation may be in the nature of the tasks that developers need to solve. In examining why developers have trouble using cryptography APIs, participants reported that they don't need cryptography very often in their daily work. They are, instead *typical application developers*, who only sometimes need cryptography [20, p.938].

Gamification features encourage developers to participate, promising status and recognition within the on-line community, two known motivators of developers in workplace environments [8] and online [21]. Links between helping behaviour and reputation among developers have been established in office-based software development environments [22] and in early investigations examining connections between willingness to contribute in on-line environments and social capital [21]. A number of other workplace motivators that might bring developers to Stack Overflow or drive them to engage have been identified, including a need for social connection, peer interaction, and identification with the task [23].

The Stack Overflow developer survey from 2016 supports these findings. In 2016, 42,134 responses were given to a question about motivation[3]. Developers indicated that they used the site to get help for tasks on the job (76.0%), but also because they love to learn (61.9%), to give help to others (46.1%) and to communicate with others "like me" (17.9%). Developers generally regard Stack Overflow as useful with answers that are of high quality [24]. However, within usable security research, it has been shown that the code samples taken from posts relating to security may not be as robust or correct as other information sources like books and vendor supplied documentation [25].

1) Asking and Answering Questions: The Stack Overflow community regulates activity on the site through extensive guideline documents and discussion about expected behaviour within pages dedicated to questions and answers about Stack Overflow operations. A developer who has a question or an answer must read a list of advice before submitting a post.

Users are encouraged to improve their posts before submitting, to search within archives before posting, and to be specific and provide details and context that will make answering easier. When asking a question, a developer must accept this advice by clicking a tick-box before seeing a page with a form. The form page offers additional advice in a "How to Ask" sidebar that notes that "we prefer questions that can be answered, not just discussed." Developers who want to learn more about asking questions can click a link that leads to a longer page with information[4]. The site also sets guidelines for how participants should behave, urging writers to ask something that is "relevant" to the larger community and asking developers to be open to suggestions or answers that are different than they anticipated[5].

2) The Role of Comments: Anyone can ask or answer a question, but to add a comment on a different user's post, a developer must have 50 reputation points. Comments are limited to 600 characters. They are described as "second-class citizens" to the question and associated answer posts that are the main sources of information on the site [6]. Comments are intended to be temporary, conceived of as "post-it notes" on the question or answer they support, that can be deleted soon after posting, or by moderators within the site [7].

In practice, many comments persist for years after they are given. Their management within the site is difficult to contain and regulate. The community debates privileges that are or should be associated with comments, why and how they should be edited, with posts dedicated within the help site to the "bad" habits of people who delete comments [8], or who answer questions within comments [9].

Stack Overflow posts develop over time, and comment streams are a part of this process. Though most edits take place soon after the posts are created, a link has been established between ongoing edits to posts and commenting activity, and between commenting activity and post edits [26]. This relationship between editing and commenting activity suggests that something about the interaction they comprise is valuable.

Interactions among Stack Overflow users that take place within comment streams and between posts and commenting streams are the area of investigation for this study.

[1]https://research.hackerrank.com/developer-skills/2018/
[2]https://insights.stackoverflow.com/survey/
[3]https://insights.stackoverflow.com/survey/2016#community
[4]https://stackoverflow.com/help/how-to-ask
[5]https://stackoverflow.com/questions/ask/advice
[6]https://meta.stackexchange.com/tags/comments/info
[7]https://meta.stackexchange.com/questions/19756/
[8]https://meta.stackexchange.com/questions/19756/
[9]https://meta.stackexchange.com/questions/4217/

III. Method

The ethnographic method is used to study peoples' actions and accounts of actions [11]. The method allows researchers to develop understanding about what practitioners working in socio-technical environments do and why they do it. The analytic stance allows researchers to consider experience from the perspective of the insider, in this case individual users of Stack Overflow.

Ethnographic research can be participatory or non-participatory. Non-participatory researchers observe people in settings but do not take part in activities [11]. It is unobtrusive, making it possible to see actions unfold as they do under normal circumstances. Stack Overflow is a naturalistic environment, a place that developers regularly use or consult in daily practice. It is also an environment in which talk is unscheduled [27], that is, not held within formal processes or to meet project constraints.

Unscheduled talk is integral to software development. Conversations between developers include stories about past experiences, but also provide narratives in the midst of practice [28] that workers use to develop confidence, and to learn [29]. Through talk, developers generate understanding of what software is and needs to be. This kind of "code talk" is often serendipitous, but lends structure to decisions about programming that will be undertaken at the desk [27].

The key in this study has been to identify features of talk about security that figure into "common-sense knowledge" [29]. To do this, principles of computer mediated discourse analysis [30] were used to isolate and catalogue features that characterise Stack Overflow as a communication medium in which messages are posted, and to examine in more detail the social or situational factors that shape interactions about security.

This study intends to strengthen investigations into the social and human aspects of software engineering, asking:

How do developers on Stack Overflow engage with one another when dealing with issues related to security?

This question relates to two research aims that are addressed in this study by examining the nature of interactions between developers.

1) To understand more about the security practices of developers who are not security specialists.
2) To understand what security "is" within the broader Stack Overflow community.

IV. Data Selection

Stack Overflow encourages participation through features that reward developers with points and badges when their posts are "voted up". Having a higher reputation grants access to different opportunities for contribution. A search of questions within the meta help site and queries in the data explorer [10] made it apparent that reputation and status are important to developers on the site. For example, the meta help site for Stack Overflow consistently lists "How does reputation work" as one of the most frequently asked questions. The data explorer shows numerous queries that users have posted to find how their reputation compares to others, or how far they are from achieving a higher status.

Threads were selected that are perceived within the Stack Overflow community to be valuable, as indicated through scoring features. Data associated with the twenty highest scored questions given the tag "security" were extracted from the hosted version of the Stack Exchange data explorer data dump of 14 January 2018. As reported in the prior study [10], data were selected using the following criteria:

1) Evident need. Top-scored questions and accepted answers were chosen to form the set. Questions indicate evidence of a need to write secure code, but with gaps in knowledge or understanding.
2) Non-specialists. To meet the guiding aim within the overarching project, data were drawn from the general Stack Overflow site rather than the specialised Information Security Stack Exchange site.
3) Stable data. Highest scored postings correspond to the list of Askers given for All Time. These posts are conducive to analysis, as they are less active than recent top rated posts.

V. Analysis

In the prior study, the set of 20 questions and 137 comments made about those questions were catalogued and given codes reflecting three broad dimensions: security advice and assessment, values and attitudes, and community involvement. Within the current analysis, the set of 20 accepted answers and 364 comments associated with the answers were catalogued to identify in more detail features of participation, and to isolate interactions relating to security. The comments were examined in two phases, described in the sections that follow.

A. Phase I

Analysis began with cataloguing to mark features of the messaging environment [31]. Commenting in Stack Overflow is asynchronous and so details of the timing of messages in relation to one another were noted, as were indications that messages were deleted, patterns of interaction within and between question and answer streams, naming and addressing techniques and quoting. This analysis also established a broad purpose for the comment, using codes developed within the preliminary study. In addition, the profile pages within Stack Overflow and Stack Exchange for each developer who answered a question and those for a subset of commenters were consulted.

A comparison of activity within comment streams for questions and answers revealed three distinct characteristics.

In contrast with findings given in the first study [10], interactions that occur within the answer comment streams were found to contain less evidence of proclamations or principles about what *should* be done in relation to security,

[10] https://data.stackexchange.com/

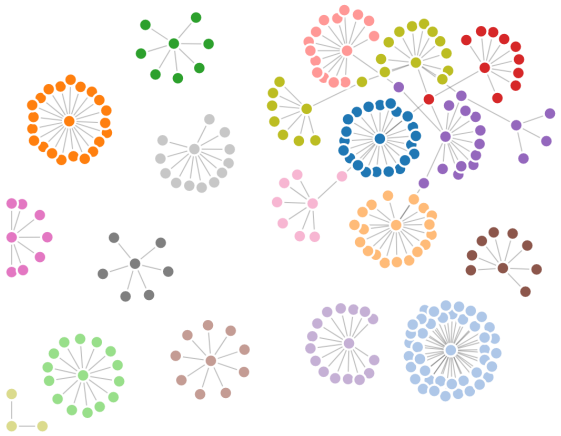

Fig. 1. This image depicts commenting activity for answers. Each answer is represented by an individual cluster. Dots around the center point of each cluster represent users who commented at least once within an answer stream. Within this set, answer streams include varying numbers of comments. Most commenting activity is isolated; few users comment on more than one stream.

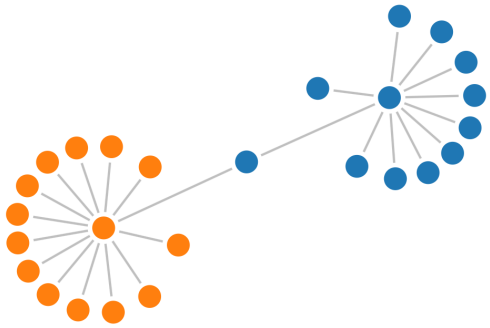

Fig. 2. This image represents commenting activity for Answer 5 and the corresponding question. The answer is depicted in blue, the question in orange. Individual dots around the center points represent commenters. The dot connected to both clusters, represents a user that commented in both the answer and question comment streams. In this case and within the set, the question asker is often the only user to comment in both streams.

and less amplification of risks and fears around security. They include more detailed information about how specific features of languages or tools work. In addition, a higher proportion of commenters within the answer stream address their comments to specific users. Within the answer stream there are also fewer indications within comments of tone or register [31] that are critical, sarcastic, or ironic. Finally, looking at commenting participation across both sets of comments, surprisingly few people were found to be active across both the question and answer streams (see Figure 1). The most likely person to comment in both streams is the Question Asker (see Figure 2).

B. Phase II

Analysis in phase II was restricted to examination of answer commenting streams. This was done in two parts. First, interactions were identified and catalogued. Next, the internal structure of messages was examined.

1) Interactions: To identify interactions, each comment was examined to identify to which stream the comment was directed and to whom it was addressed. Pairs and exchanges identified were using evidence that users negotiated turn-taking and maintained cross-turn coherence [30] within individual posts. Coherent interactions were indicated when:

1) Participants consistently addressed or quoted each other in their comments
2) Comments persisted: there was no evidence of deletion, and the users who created the comments remain members of Stack Overflow, and
3) Comments were adjacent within streams, posted at close intervals in time to one another, or used public names or quoting after time passed to unambiguously indicate a relationship to a prior comment.

Interactions of four kinds were identified within the answer comment streams.

1) Pairs. Most interactions form pairs between two individual users rather than in extended exchanges with one or more users.
2) Three-part exchange. People who left two comments frequently participated in a single three-part exchange with one other person. Exchanges include an initiating comment, a response, and a follow-up comment that confirms understanding, provide thanks, to apologise or retract a statement, or otherwise close the interaction in some way.
3) Multi-part exchange. Often between two people, characteristic exchanges of this type include challenges or a series of questions and responses.
4) Broadcasts. Within broadcasts, multiple developers chime in on a single topic. In both cases within this set, broadcasts are used to situate the security problem within time, indicating how companies handled license key generation in the past (Q8) or noting browser updates over time (Q19).

2) Structure and Purpose: Assigning a single code to indicate purpose or intent to comments is, in many cases, not possible. Within a single comment, developers often convey more than one piece of information. They might offer a suggestion for an alternative solution, while at the same time indicating that they are not confident, and need help.

Many comments are similar, and reflect patterns of moves or schema identified in other studies examining asynchronous communication. For example, messages sent to academic mailing lists have been found to commonly follow three moves: a reference is made to an earlier message, a view is expressed, and an appeal is made to other participants [32]. Messages in the set examined here also often have a similar structure. Moves were examined to understand the language developers

TABLE I
ACCEPTED ANSWERS FOR 20 TOP-SCORED QUESTIONS, 14 JANUARY 2018.

	Asked	Answered	Accepted	Question Comments	Answer Comments	Tags
A1	13.12.12	13.12.12	17.12.12	27	8	android; proguard;reverse-engineering
A2	16.1.12	16.1.12	16.1.12	12	26	java;string; passwords;char
A3	26.7.11	26.7.11	27.7.11	0	19	hash;internals;bcrypt+J6
A4	2.7.11	2.7.11	2.7.11	1	2	authorization;authentication
A5	21.4.11	21.4.11	31.7.13	17	14	php;mysql;sql; sql-injection
A6	9.2.11	9.2.11	10.2.11	6	13	encryption;hash;cryptography
A7	15.8.10	26.8.11	17.9.12	0	12	oauth;access-token;refresh-token
A8	8.6.10	16.6.10	5.11.10	9	17	cryptography
A9	19.4.10	19.4.10	19.4.10	5	50	javascript;json;ajax
A10	17.2.10	28.2.10	1.3.10	18	25	password-encryption;password-storage
A11	4.2.09	4.2.09	10.9.15	0	7	windows
A12	30.12.08	30.12.08	30.12.08	7	49	php; passwords;hash;protection
A13	1.12.08	1.12.08	15.3.15	9	10	validation;sql-injection
A14	9.10.08	9.10.08	6.6.09	1	4	post;encryption;https;get
A15	25.9.08	25.9.08		1	17	php; pdo;sql-injection
A16	24.9.08	24.9.08	24.9.08	9	27	php; xss;sql-injection;user-input
A17	18.9.08	18.9.08	19.9.08	3	17	php;database
A18	17.9.08	17.9.08	17.9.08	5	15	javascript;performance; eval
A19	28.8.08	28.8.08	28.8.08	6	15	browser;autocomplete;passwords
A20	11.8.08	11.8.08	11.8.08	1	17	wcf; rest;authorization;rest-security

use when they initiate and respond within interactions, and to identify kinds of information given in responses.

The analysis identified words and punctuation that signal tonal features of messages as well as indications of information trading about security techniques, scenarios, circumstances and principles. Within the paired interactions, the kinds of things developers asked were found to have commonalities with other studies [17], [33]. In general, users asked:

- how security concerns relate to individual circumstances
- for more information, including different sources
- for technical help

However, because this analysis focuses on both sides of the interaction, it was also possible to establish how developers respond to questions. Responses fall into the following broad categories:

- explanations of how technologies work
- establishing security facts
- confirming that understanding is correct
- assessing how alternative language features, tools or frameworks apply to the security issue under discussion

VI. FINDINGS

This section provides a structured look at how the community of Stack Overflow operates in answer comment threads that have a security focus, and includes a qualitative examination of how developers within the threads describe security to one another, how they display understanding about security, and the way they apply secure practices to programming tasks.

A. Posts

The list of 20 accepted answers is summarised in Table I. The table indicates dates for question and answer posts, the date the answer was accepted, and the number of comments for the question and answer streams. The tags are those associated with the question; *security* was removed.

The questions were asked between August, 2008 (Q20) and December, 2012 (Q1). 17 of the questions remained active in 2017 and 2018. All questions except three (Questions 3, 7 and 11) had at least one comment given about the question. Many of the answers were accepted within one month of being given; however, some were not accepted until years after being asked. This discrepancy in dates may reflect that answers can lose or gain accepted status within the community over time. The set includes issues that are several years old.

B. Participants

As previously reported, twenty different users asked questions. Six of the question askers participated in the question comment stream; a few askers also participated in the answer comment stream. With one exception, these developers do not engage in discussion about the answer, but use comments to give thanks or feedback about the quality of the answer, or to provide detail about technologies or techniques that are in use. The accepted answers were likewise provided by twenty distinct users of Stack Overflow. The answers are highly rated within the site and three have received a bounty, a reputation award given to answers by the askers. Fifteen of these users participate within the comment streams for their answer, but only five commented six or more times. Surprisingly, none of the users who submitted answers comment within the question stream for their own question, though a few answer in streams for other questions (see Table II.

Though half of the answer providers are members of the Information Security community, only six have been recently active. Their activity within posts tagged with security varies; four of the answerers appear in the Stack Overflow list of top twenty security question answerers of all time, and several of the users frequently participate in threads tagged with security. Taken together, activity within this group suggests that, as with the askers of questions the developers giving answers

are primarily non-specialists who exhibit a range of levels of activity within the security channel and the Information Security Stack Exchange site.

A slightly greater sense of security related activity can be seen by looking at an overview of information for Answer providers drawn from the wider site. Only one answer provider, EpicRainbow, identifies within their profile description as having an interest or expertise in security. However, for half of the answer providers, the top 3 highly voted tags associated with the answerers suggest that other Stack Overflow users recognise and regard participation these developers make in posts that include security as a tag.

C. Answer Comment Streams

Within the answer streams, 250 Stack Overflow users made 364 comments. The majority of users, 197 left a single comment, 32 left two comments, 10 left three comments, and 11 left more than three comments.

Comment streams for questions and answers are distinct. A high proportion of commenters within the answer stream address their comments to specific users, either by referencing the user's public name, through direct quoting or referencing concepts given within a prior answer. However, many comments are directed toward the writer of the answer post. In these cases, direct addressing is not used, but the comment may include quotes of the answer, or clear references to concepts within the answer. Within the answer stream there are also fewer indications within comments of tone or register [31] that are critical, sarcastic, or ironic.

D. A Worked Example

Following is a representative set of comments given by three users within the answer stream for Question 5. The commenters are Nemo, Smee and JohnnyGianni. Each of these writers left only one comment in this answer stream. Nemo has participated the least in the security channel, with participation in only 11 question or answer posts. Though Smee is active in the security channel, having participated in 135 posts, there is no indication given within profile information of interest or expertise in security. By contrast, JohnnyGianni is less active in the security channel (41 posts), but more active in the Information Security site and makes reference to security experience within the profile description.

Each comment is used to illustrate aspects of interaction across the larger set of comments in the answer streams. In these extracts, different moves [32] are segmented (eg. S1, S2). Information is given in brackets ([]) following each segment to indicate a code given during analysis to indicate a purpose for the segment within the comment. Almost a month passes between the first comment (A5.C3) and the second. There is a seven-month gap in time between the second comment (A5.C4) and the response (A5.C5).

Answer 5, Comment 3: Nemo

```
S1 vintage [direct address A5]
S2 `$iId =
   mysql_real_escape_string((int)"1;
   DROP table");` [technique]
S3 or `$dirty = "1; DROP table";$iId=
   mysql_real_escape_string((int)$dirty);
   [technique]
S4 would be a better example than
   what you have [view]
S5 I think [judgment]
S6 Nemo 09/09/2011 05:09 [A5.C3]
```

Often paired interactions in the corpus are initiated in reference to information given in the answer post, as comment Answer 5, Comment three above demonstrates. Nemo is critical of the accepted answer given by Manfred for Question 5 (S5) but only provides an alternative solution within a comment, not within an answer post. This type of answering is a recognised behaviour within the community.

The comment given in Answer 5, Comment 4 by Smee is initiated in response to a comment made earlier in the comment stream for Answer 5. Because the original commenter (Nemo) does not respond, the comments A5.C4 and A5.C5 have been treated in analysis as a paired interaction.

Answer 5, Comment 4: Smee

```
S7 But this [ref A5.C3]
S8 wouldn't be a real problem, [view]
S9 because `mysql_query()` doesn't execute
    multiple statements, [proof]
S10 no?"  [appeal]
S11 Smee  07/10/2011 21:07 [A5.C4]
```

Smee challenges the alternative example suggested by Nemo, but indicates with the phrasing (S10) and use of a question mark that he is not certain about the proof that is given. The appeal he makes ("no?") invites a response.

Answer 5, Comment 5: JohnnyGianni

```
S12 Smee [direct address]
S13 Although the usual example is
    `DROP TABLE` [ref A5.C3]
S14 in practice the attacker
     is more likely [scenario]
S15 to `SELECT passwd FROM
     users`.[technique]
In the latter case,
S16 the second query is usually
    executed by use of a
    `UNION` clause."
    [technique]
S17 JohnnyGianni 21/05/2012 09:47 [A5.C5]
```

Nemo does not reply to Smee. The comment made by JohnnyGianni, given several months later, contains information about how SQL can be applied in a particular kind of security attack. The comment also makes an assessment of the quality of sources of security information that are available. The

TABLE II
ACCEPTED ANSWER AUTHORS. ASTERISKS (*) INDICATE PARTICIPATION IN A COMMENT STREAM FOR A DIFFERENT QUESTION AND RECENT ACTIVITY IN THE INFO SECURITY SITE.

Code	Pseudonym	Answered	Q Comment	A Comment	Info Sec	Posts w Security	Top Tags for User (by Vote)
A1	ExperiencedPigeon72	13.12.12	n	y	y	2	android; security; reverse-engineering
A2	**hercules**	16.1.12	n	y	n	51	c#; java; .net
A3	**FortuneRat**	26.7.11	n	y	y*	146	java; security; encryption
A4	recipegod	2.7.11	n	y	y*	19	c++; c++11; c
A5	vintage	21.4.11	n	n	n	7	php; mysql; sql
A6	**EpicRainbow**	9.2.11	n	y	y	143	php; security; mysql
A7	Techniq	26.8.11	n	n	y*	2	security; refresh-token; access-token
A8	HeroJan	16.6.10	n*	y	y*	17	algorithm; c#; sql
A9	newton	19.4.10	n	n	n	2	javascript; ajax; security
A10	Syntaxis	28.2.10	n	y	y*	18	c++; c; c#
A11	rabbitsfoot	4.2.09	n	y	n	4	windows; security; c#
A12	**Anthropic**	30.12.08	n	y	n	5	php;hash;security
A13	mutator	1.12.08	n	y	n	14	c#; .net; wpf
A14	Lemongrass	9.10.08	n	n	n	18	javascript; function; syntax
A15	ColMustard	25.9.08	n*	y	y*	47	c#; .net; sql
A16	**darth**	24.9.08	n	y	n	15	php; security; sql-injection
A17	whatever	18.9.08	n	y	n	2	algorithm; language-agnostic; mergesort
A18	Einstein	17.9.08	n	y	n	1	javascript; hex; tostring
A19	codfish109	28.8.08	n	n	y	4	c#; .net; datetime
A20	candyfunctions	11.8.08	n	y	y	37	git; python; c++

comment notes that how attacks are "usually" described is different from the techniques used by attackers with SQL"in practice". Finally, information is included about the structured query language that is phrased in neutral terms. The last line (Segment 16) might be associated with attacking activity, but can also be read as a correction or lesson for Smee about how the structured query language works.

E. Developing Awareness and Knowledge

Interactions within comment streams for answers in Stack Overflow support the development of security awareness and knowledge in three ways:

1) **Provide focused assistance.** Interactions provide developers with information, clarification or corrections and confirm understanding. Often this kind of support is freely given without indications of judgment or criticism.
2) **Associate technology facts with security problems.** This linking is often material, for example in associating small details about how a language works with an equally small feature of security. Smee was correct, the function doesn't execute multiple statements, but JohnnyGianni explains that there are other ways to use the query language (S16) that will give similar results.
3) **Situate advice in the security landscape.** Many responses situate the advice given within the larger sphere of security discourse. This is often done with subtle language cues, as in JohnnyGianni's indication to Smee that the way an attack is usually conveyed does not match what is done in practice. At other times, the alert more directly situates technical information within the broader security landscape, for example by explaining attack scenarios.

F. Characteristics of Engagement

The worked example shows three comments made for a single answer answer. In this example, the three commenters left only one comment each in this answer stream. While this example is representative of many of the interactions in the set, there are four other characteristics of participation that should be noted.

1) **Security is complex**. Developers indicate that they recognize that security is a complex concern, and one for which information is vague, contradictory, or sparse. As one developer put it, *"I'm still learning here. ...every time I read something that makes sense, I soon notice 5 other posts that contradict it. that round-and-round gets dizzying quickly :)" (A12.C11)*

   Participants in the examined threads comment and agree on this point. However, they also counter it when they can, by suggesting sources that might be useful. As Anthropic (Question 12) offered in reply to the commenter above, *"Absolutely! I've just shared what I've found. I found a number of things from Shneier on Security and a very long (convoluted) discussion on a news site (don't remember which now)." (A12.C12)*
2) **Support**. There is evidence of support for answers given in the form of verbal comments that indicate things like "Great answer", "Thanks for the concise answer", or "I never would have thought of that.". In the case of the worked answer, Epic Rainbow commented elsewhere in the stream to support the answer given to Answer 5 by Manfred, having perceived that down-voting activity by the community in this case was unfair:

   "To the people downvoting this answer: this answer is completely correct. This is far more likely to be

the reason your use of mysql_real_escape_string is going to be compromised than my answer below. This belongs as the accepted answer (but both can live together)..." (A5.C8)

3) **Derision**. The acceptance of answers is, at times, contentious. Developers note that answers are not correct, address the wrong topic, or are out of date.
   There are several instances of a user challenging a detail within the accepted answer, and then using the opportunity to draw attention to the answer he or she has written. These reveal workings of Stack Overflow as a community, and demonstrate that users participate for many reasons. In the prior example, EpicRainbow was supportive, however, was dismissive of the answer given to Question 15, commenting *"See my answer below for a demonstration and explanation of an attack..."* (A15.C5).
   This kind of community level activity has an impact within the site. The answer analysed for Question 15 has, in the months since data capture in January 2018, lost accepted answer designation (Table I). The comment stream for the answer suggests that the answer was not accepted by community almost from the moment it was posted, with comments like EpicRainbow's, that use a negative or teasing tone or iconography.
4) **Passing time**. Answers are changed based on comments made in the answer stream. Generally, users that answer question note this in the body of the answer, using text like "As noted in the comments" or more directly recognising the contributions of particular users: "Edited as per Joe's astute comment".
   The list includes issues that are several years old, making it possible to explore features of community development across a longer span of time. For example, reference is made in one comment stream to the "brand new" Information Security channel. The reason for a bump in activity for another thread is noted to be an early tweet made by one of Stack Overflow's founders. The threads also give a sense of the changing relevance and importance of particular issues at different points in time. Commenters use streams to note when information is out of date or to broadcast up-to-date information.

VII. Discussion

Stack Overflow exhibits many attributes of a community of practice [19]. Many of the features of interaction identified in this analysis relate or reflect the Stack Overflow community as a whole.

Activity within Stack Overflow centres around the domain of programming, and the collective need developers have to learn. It is fair to say that over the course of a decade, the collective process of asking and answering questions about programming has bound an international community of developers together, and that interactions within the site have an effect on software development practice in offices, schools and other environments around the world.

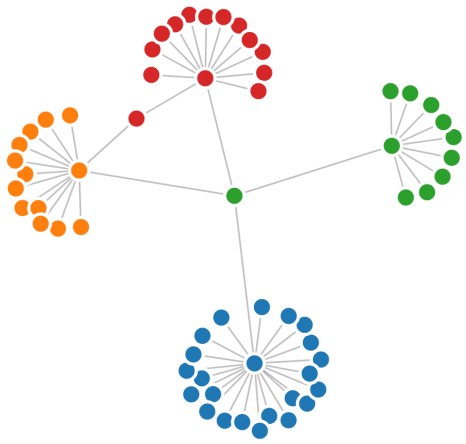

Fig. 3. This image represents commenting activity within answer streams for EpicRainbow, the answer provider for Question 6. EpicRainbow is the green dot in the middle. Within these streams, ER made six comments on four answers depicted in orange, red, green and blue.

Where, then, does security sit within Stack Overflow? It is an active topic within the site, but it is difficult to make a case from these findings that the participants in this set form a community around the practice of security.

A. Connecting with Security

The developers on Stack Overflow have a collective need to learn about security, which must be reconciled with and applied to specific programming tasks. The participation patterns within and between comment streams suggest that security is supported by a *network of practice* built through personal interactions. It is through these connections that developers guide one another, sharing information and giving help about security. [13].

Activity is not driven by answers given by few participants who are security specialists or who frequently respond to posts across different answer streams. Though there are a few answer providers and commenters like EpicRainbow who contribute more frequently across the set (see figure 3) or Anthropic who clearly exhibits advanced security understanding about cryptography in the post for Answer 12 (see also Table II), there is not the sense that these users alone hold the security channel together.

There is significant evidence within the set of ongoing editorial activity. In some circumstances, different users take over editing and updating of answers. This analysis, in-line with other studies [26], indicates that curatorial edits to answer posts are, at times, followed by increased commenting activity. However, the value and vitality of the posts does not come out of community-level activity to improve grammar or to keep links up-to-date. Instead, the significance lies in the impact that interactions within comment streams have on answer posts. Answers are updated to reflect changes in tools or techniques that can be used to address a security problem, but also

changes in thinking about what is significant to represent for a given solution.

B. Tending to Security

Security has been described as a secondary concern to developers, one that is prioritized alongside other tasks developers need to complete [34]. The threads in this analysis, oriented as they often are to language features or use of APIs, support this claim, with some caveats. This analysis demonstrates that the network of developers connecting within Stack Overflow *tend to* the problem of security within the context of the technical solutions that are given by answer providers.

Tending is easiest to see within the commenting patterns of question answerers. The most frequent commenters in the set are users who answered questions (see names in bold within Table II). Their reliable presence makes a difference to the coherence of the thread. It is easier to follow the development of the issue over time when it is anchored by back and forth between an answer provider and different commenters. These developers may also be known and trusted within the network or larger community of Stack Overflow, factors that have been associated with security tool adoption [35].

Tending is also apparent in the links that are established between the security concerns in a question, and the task at hand. Arguments draw in concepts and points made by other commenters, reference other streams, and refer to sources outside of Stack Overflow. These sources include blog posts and news items, but also draw heavily on examples of how security is handled in other technologies and languages.

On-line sources of guidance about secure software development have been found to contain gaps in coverage, and developers need to rely on diverse sources of information [36]. Findings in this study suggest that developers are aware of this, but also comfortable drawing upon various sources to build understanding.

Research has suggested that "challenge" talk between developers, rather than formal processes or artefacts, is the best way to develop techniques for security among developers [37]. Developers in this set do challenge each other, but do not, in the main, identify themselves in comments as upholding security or protecting code from attackers. Things that an attacker might do with code or in languages are conveyed as part of programming, described in terms of techniques that developers might also use in code and with languages.

Given the opportunity, it has been shown that developers turn to Stack Overflow to find solutions to security problems, however the code samples taken from security posts may not be as robust or correct as other information sources like books and vendor supplied documentation [34] The analysis performed here cannot comment on the quality of the information supplied, however the threads make it clear that developers do not blindly accept the information they are given.

Instead, the evidence shows that developers ask for more information, and ask related questions. It is also clear that developers correct each other, explaining how technologies work, but also illuminating security implications in specific situations. Finally, developers respond to one another over long stretches of time. Points made by commenters result in edits to the answers for many years. The breadth of queries and comments over time suggests that developers continue to require and lend support to one another in understanding the significance of security in relation to particular tasks, technologies and tools.

VIII. Limitations

There is an inherent bias in approaching security by looking at discussion in a programming environment. The mandate of the site is to help developers solve programming problems, and so discussion naturally centres within the answer streams around technical aspects of software development.

There are also other limitations in the sampling process used to gather data and in the quality of the data set used in analysis. The top twenty list has shifted by one since data were collected. Users can change their identity, which makes it hard to link comments to one another. Users can also leave Stack Overflow. In these cases, the comments still appear on the website, but must be explicitly requested in queries for data. Finally, there are gaps in the record, with clear evidence that comments have been deleted. There are also a small number of the answers posted in 2008 or 2009 for which comments are unavailable before dates in 2012. To counter these limitations in the data, analysis has focused on isolating interactions within threads, rather than conceiving of the threads as conversations in their entirety and through careful examination of addressing and quoting techniques and cross-examination of streams to establish relationships between users and identities.

IX. Conclusions

Secure coding practice is supported on Stack Overflow through a network of interactions. Users guide and engage with one another through conversations that:

- Provide focused, individualized assistance.
- Associate technology facts with security problems.
- Situate advice in the security landscape.
- Broadcast details or facts that orient the security issue in time.

Developers actively foster security knowledge on the site by writing and developing answer posts, and by dropping in on comment streams to share information and receive help from one another. The developers who take part in posts see security as something interesting to pursue and to tinker with. As a forum for exchanging information, Stack Overflow is a relevant resource, allowing developers to thoughtfully connect with and tend to security problems.

Acknowledgment

Supported by the National Cyber Security Centre (NCSC). Nuseibeh thanks SFI, EPSRC and ERC for financial support.

REFERENCES

[1] "Cyber Resilience: Playbook for Public- Private Collaboration," World Economic Forum, Tech. Rep. [Online]. Available: http://reports.weforum.org/cyber-resilience/

[2] H. Berghel, "Equifax and the Latest Round of Identity Theft Roulette," *Computer*, vol. 50, no. 12, pp. 72–76, Dec. 2017.

[3] C. Stewart III, "The 2016 US Election: Fears and Facts About Electoral Integrity," *Journal of Democracy*, vol. 28, no. 2, pp. 50–62, 2017.

[4] S. Furnell and A. Rajendran, "Understanding the influences on information security behaviour," *Computer Fraud & Security*, vol. 2012, no. 3, pp. 12–15, Mar. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361372312700532

[5] C. Posey, T. L. Roberts, P. B. Lowry, and R. T. Hightower, "Bridging the divide: A qualitative comparison of information security thought patterns between information security professionals and ordinary organizational insiders," *Information & Management*, vol. 51, no. 5, pp. 551–567, Jul. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378720614000421

[6] S. Furnell and K.-L. Thomson, "From culture to disobedience: Recognising the varying user acceptance of IT security," *Computer Fraud & Security*, vol. 2009, no. 2, pp. 5–10, Feb. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361372309700193

[7] I. Becker, S. Parkin, and M. A. Sasse, "Finding Security Champions in Blends of Organisational Culture," 2017.

[8] S. Beecham, N. Baddoo, T. Hall, H. Robinson, and H. Sharp, "Motivation in Software Engineering: A systematic literature review," *Information and software technology*, vol. 50, no. 9, pp. 860–878, 2008.

[9] H. Sharp, H. Robinson, and M. Woodman, "Software engineering: community and culture," *IEEE Software*, vol. 17, no. 1, pp. 40–47, Jan. 2000.

[10] T. Lopez, T. T. Tun, A. Bandara, M. Levine, B. Nuseibeh, and H. Sharp, "An Investigation of Security Conversations in Stack Overflow: Perceptions of Security and Community Involvement," in *Proceedings of the 1st International Workshop on Security Awareness from Design to Deployment*, ser. SEAD '18. New York, NY, USA: ACM, 2018, pp. 26–32.

[11] H. Sharp, Y. Dittrich, and C. R. B. d. Souza, "The Role of Ethnographic Studies in Empirical Software Engineering," *IEEE Transactions on Software Engineering*, vol. 42, no. 8, pp. 786–804, Aug. 2016.

[12] J. Atwood, "Introducing Stackoverflow.com," Apr. 2008. [Online]. Available: https://blog.codinghorror.com/introducing-stackoverflow-com/

[13] E. Wenger, B. Traynor, and M. de Laat, "Promoting and assessing value creation in communities and networks: a conceptual framework," Open University of the Netherlands, Ruud de Moor Centrum, Tech. Rep. Rapport 18, 2011.

[14] M.-A. Storey, L. Singer, B. Cleary, F. Figueira Filho, and A. Zagalsky, "The (r) evolution of social media in software engineering," in *Proceedings of the on Future of Software Engineering*. ACM, 2014, pp. 100–116.

[15] B. Vasilescu, A. Serebrenik, P. Devanbu, and V. Filkov, "How Social Q&A Sites Are Changing Knowledge Sharing in Open Source Software Communities," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ser. CSCW '14. New York, NY, USA: ACM, 2014, pp. 342–354. [Online]. Available: http://doi.acm.org/10.1145/2531602.2531659

[16] A. Zagalsky, C. G. Teshima, D. M. German, M.-A. Storey, and G. Poo-Caamao, "How the R community creates and curates knowledge: a comparative study of stack overflow and mailing lists," in *Proceedings of the 13th International Conference on Mining Software Repositories*. ACM, 2016, pp. 441–451.

[17] S. M. Nasehi, J. Sillito, F. Maurer, and C. Burns, "What makes a good code example?: A study of programming Q&A in StackOverflow," in *2012 28th IEEE International Conference on Software Maintenance (ICSM)*, 2012, pp. 25–34. [Online]. Available: doi.ieeecomputersociety.org/10.1109/ICSM.2012.6405249

[18] R. Slag, M. d. Waard, and A. Bacchelli, "One-Day Flies on StackOverflow - Why the Vast Majority of StackOverflow Users Only Posts Once," in *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*. Florence, Italy: IEEE, May 2015, pp. 458–461.

[19] E. Wenger, R. A. McDermott, and W. Snyder, *Cultivating Communities of Practice: A Guide to Managing Knowledge*. Harvard Business Press, 2002, google-Books-ID: m1xZuNq9RygC.

[20] S. Nadi, S. Krger, M. Mezini, and E. Bodden, ""Jumping Through Hoops": Why do Java Developers Struggle with Cryptography APIs?" in *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)*, May 2016, pp. 935–946.

[21] M. M. Wasko and S. Faraj, "Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice," *MIS Quarterly*, vol. 29, no. 1, pp. 35–57, 2005. [Online]. Available: https://www.jstor.org/stable/25148667

[22] A. Hargadon and B. A. Bechky, "The politics of knowledge work in a software development group," *Qualitative organizational research*, pp. 15–35, 2005.

[23] H. Sharp, N. Baddoo, S. Beecham, T. Hall, and H. Robinson, "Models of motivation in software engineering," *Information and software technology*, vol. 51, no. 1, pp. 219–233, 2009.

[24] M.-A. Storey, A. Zagalsky, F. F. Filho, L. Singer, and D. M. German, "How Social and Communication Channels Shape and Challenge a Participatory Culture in Software Development," *IEEE Transactions on Software Engineering*, vol. 43, no. 2, pp. 185–204, Feb. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7498605/

[25] Y. Acar, S. Fahl, and M. L. Mazurek, "You are not your developer, either: A research agenda for usable security and privacy research beyond end users," in *Cybersecurity Development (SecDev), IEEE*. IEEE, 2016, pp. 3–8. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7839782/

[26] S. Baltes, L. Dumani, C. Treude, and S. Diehl, "SOTorrent: Reconstructing and Analyzing the Evolution of Stack Overflow Posts," *arXiv:1803.07311 [cs]*, Mar. 2018, arXiv: 1803.07311. [Online]. Available: http://arxiv.org/abs/1803.07311

[27] A. Higgins, "'Code talk' in soft work," *Ethnography*, vol. 8, no. 4, pp. 467–484, Dec. 2007. [Online]. Available: https://doi.org/10.1177/1466138107083563

[28] J. E. Orr, "Narratives at work: Story telling as cooperative diagnostic activity," in *Proceedings of the 1986 ACM conference on Computer-supported cooperative work*. ACM, 1986, pp. 62–72.

[29] P. Duguid, "What talking about machines tells us," *Organization Studies*, vol. 27, no. 12, pp. 1794–1804, 2006.

[30] S. C. Herring and J. Androutsopoulos, "Computer-mediated discourse 2.0," *The handbook of discourse analysis*, vol. 2, pp. 127–151, 2015.

[31] S. C. Herring, "A faceted classification scheme for computer-mediated discourse," *Language@ internet*, vol. 4, no. 1, 2007.

[32] S. C. Herring, S. Barab, R. Kling, and J. Gray, "An approach to researching online behavior," *Designing for virtual communities in the service of learning*, vol. 338, 2004.

[33] C. Treude, O. Barzilay, and M.-A. Storey, "How do programmers ask and answer questions on the web?: Nier track," in *Software Engineering (ICSE), 2011 33rd International Conference on*. IEEE, 2011, pp. 804–807.

[34] Y. Acar, M. Backes, S. Fahl, D. Kim, M. L. Mazurek, and C. Stransky, "You Get Where You're Looking For: The Impact Of Information Sources on Code Security," in *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 2016, pp. 289–305.

[35] S. Xiao, J. Witschey, and E. Murphy-Hill, "Social influences on secure development tool adoption: why security tools spread," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 2014, pp. 1095–1106.

[36] Y. Acar, C. Stransky, D. Wermke, C. Weir, M. L. Mazurek, and S. Fahl, "Developers Need Support, Too: A Survey of Security Advice for Software Developers," in *Cybersecurity Development (SecDev), 2017 IEEE*. IEEE, 2017, pp. 22–26.

[37] C. Weir, A. Rashid, and J. Noble, "I'd Like to Have an Argument, Please: Using Dialectic for Effective App Security," 2017.